Counterfactual de se*

Hazel Pearson
Queen Mary University of London

Abstract This paper addresses a long-standing debate concerning the derivation of de se construals. One camp holds that there is a dedicated mechanism of 'de se binding', which results in a de se pronoun being interpreted as a variable ranging over the doxastic alternatives of the attitude holder (e.g. Chierchia 1990) . Another treats de se as a special case of de re under the acquaintance relation of identity (e.g. Lewis 1979, Reinhart 1990) . This debate is premised on the assumption that the two different routes to de se result in identical truth conditions. I argue that this assumption is incorrect for a class of cases that can be delineated in a principled fashion - counterfactual attitude reports involving counter-identity, such as *Ivanka imagined that she was Melania and she was giving an interview*. The argument builds on Ninan 2008, who noticed that de re construal works differently with counterfactual attitudes, and that this has consequences for de se interpretation in this type of sentence. I spell out these consequences more precisely, drawing on a novel, crosslinguistically robust generalization about unambiguously de se expressions such as PRO (the 'De Se Generalization'). I argue that a treatment of such expressions that appeals to de se-as-de re cannot account for the De Se Generalization in a principled way, and hence that a dedicated mechanism of de se binding must be included among the expressive resources of the grammar.

Keywords: attitude reports, de se, de re, counterfactual attitudes, obligatory control, indexical shift

* For helpful comments and discussion, I thank Gennaro Chierchia, Jeruen Dery, Vera Hohaus, Hans Kamp, Natasha Korotkova, Emar Maier, Andreea Nicolae, Dilip Ninan, Orin Percus, Konstantin Sachs, Paolo Santorio, Frank Sode, Stephanie Solt, Yasutada Sudo and Igor Yanovich. Thanks also to Dorothy Ahn and Yangsook Park for Korean judgments, and to audiences at ZAS Berlin, the SynSem lunch seminar at the University of Oslo, Queen Mary University of London, the New York Philosophy of Language Workshop, the University of Göttingen and the NonFinite Subjects Conference at the University of Nantes. Special thanks to *S&P* editor Josh Dever, Pranav Anand and two further anonymous reviewers for detailed comments on the manuscript which led to a much improved paper, and to Mohammad Alhailawani for typesetting assistance. The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP72007- 2013) under grant agreement no. 618871.

1 Introduction

Here is a story that was told to me by a friend who wishes to remain anonymous.

(1) One morning I was sitting with friends in the local McDonald's, as was our custom, prior to heading up the hill to school. The corner in which we frequently perched ourselves – a spot, enclosed on two sides by wall-high glass, that allowed us to gaze at the antics of pedestrians outside – this corner was today already taken, and so we grumpily found another table to congregate around nearer the middle of the establishment. While my friends were talking amongst themselves, I found myself momentarily distracted by a man sitting a short distance away, staring roughly back at me. For reasons I can't define, I took an immediate dislike to the appearance of this man, his face and countenance, and came to the swift judgment that he was some form of loathsome idiot. It was startling, therefore, a short moment later, to realize that I was gazing into a wall-high mirror – and had thus cast such an aspersion upon myself.

In this story, there is a certain individual whom our friend – call him Jonathan – believes to be a loathsome idiot (at least up until the moment when he realizes that he is looking in a mirror). Who is that individual? We know, even though Jonathan has not yet figured it out, that it is Jonathan himself. So can we then describe what happened with the following sentence?[1]

(2) Jonathan$_i$ believed that he$_i$ was a loathsome idiot.

Yes and no. On the one hand, it seems that we can coherently utter (3).

(3) Jonathan$_i$ believed that he$_i$ was a loathsome idiot, although he didn't realize that he was the person that he had in mind.

But we could also say:

(4) Jonathan$_i$ didn't believe that he$_i$ was a loathsome idiot, because he didn't realize that he was the person that he had in mind.

In the second discourse, the fact that Jonathan didn't realize that the person he was ascribing loathsome idiocy to was himself is grounds to reject (2); in (3), (2) is taken to be true *despite* this fact. It seems then that (2) has two interpretations: one on which Jonathan's failure to recognize himself is relevant to the truth value of the

1 I use coindexing merely to indicate intended coreference; the indices are not meant to have theoretical status.

sentence, and one on which it is not. On the former reading, the sentence is judged false, and on the latter one it is judged true.

The difference seems to come down to the interpretation of the pronoun: on the former reading, we shall say that the pronoun is read *de se*, while on the latter one it is read *de re*. A de se report of an attitude (a belief, desire, speech act, etc.) is a report of an attitude that in some intuitive sense is 'about' the attitude holder, and which furthermore the attitude holder is aware is about him- or herself. Jonathan's belief that the person he is looking at is a loathsome idiot has the first property, since the person Jonathan is looking at is none other than Jonathan himself. However it lacks the second property, since Jonathan does not know that he is looking in a mirror. Attitudes that have both properties – that is, that satisfy both the 'aboutness' condition and the 'awareness' condition – are attitudes *de se*. If a pronoun is construed de se, then the attitude report in which it occurs is only true if both conditions are satisfied, and the reported attitude is therefore an attitude de se . Hence (2) is judged false on the de se construal: Jonathan's belief of himself that he is a loathsome idiot does not satisfy the awareness condition. On the other hand, de re beliefs that are about the attitude holder are not subject to the awareness condition; hence if the pronoun is construed de re, then we judge the sentence true.

The topic of this paper is the proper analysis of de se reports. Beginning with Chierchia 1990, some researchers have argued that there is a dedicated mechanism of 'de se binding' that produces de se truth conditions (e.g. Anand & Nevins 2004, Percus & Sauerland 2003a,b, von Stechow 2002, 2003). On this view, the ambiguity of (2) comes about because the pronoun can, but does not have to be de se bound. On the other hand, Lewis 1979 noticed that de se construal could be derived via de re construal of the pronoun under a special acquaintance relation which we might call 'SELF'. Reinhart 1990 and subsequently (Maier 2006, 2009, 2011) argued that this 'de se via de re' route is sufficient: de se construal can be derived without appeal to de se binding, which therefore should not be postulated.

Throughout this debate, it has been taken for granted that the two 'routes to de se' result in equivalent truth conditions: the interpretation that de se binding produces is identical to that which would result from de re construal under the SELF acquaintance relation. I shall argue that this assumption is incorrect for a particular class of environment – namely reports of an attitude holder putting herself in the shoes of someone other than who she actually is. Such cases arise with counterfactual attitude verbs such as *imagine*, *wish*, *pretend*, *say* and *claim* as in the following examples:[2]

2 By 'counterfactual attitude verb' I mean a verb denoting an attitude that one can hold coherently towards a content while simultaneously believing that that content is false.

(5) a. Sophia is imagining that she is Michelle Obama and she is married to Barack Obama.

b. Sophia is wishing that she was Michelle Obama and she was married to Barack Obama.

c. Sophia is pretending that she is Michelle Obama and she is married to Barack Obama.

d. Sophia says/claims that she is Michelle Obama and she is married to Barack Obama.

On the most salient construal of the second pronoun in (5a), *Sophia* is not imagining that Sophia is married to Barack Obama. This would require her to imagine that Barack Obama has different properties than he actually has with respect to whom he is married to. Rather, the counterfactual worlds that Sophia is entertaining resemble the actual world (and resemble the worlds that Sophia's belief state designates as candidates for the actual world) in as much as in those worlds too, Barack Obama is married to Michelle Obama. What is different is that according to what Sophia imagines, she *is* Michelle Obama. What it could mean for Sophia to be Michelle Obama is a vexed topic – surely there are no worlds in which this is the case? – and we shall be returning to it shortly. For now, I shall just assume that (5a) reports an event involving Sophia putting herself into Michelle Obama's shoes, say because she is wondering what it would be like to be her. This captures our intuitions about the meaning of the sentence well enough for now. I shall argue that this interpretation is best captured by appealing to de se binding, not de se as a special case of de re. Therefore, there is de se binding.

The paper is structured as follows. Section 2 describes the two approaches to de se construal that are the focus of this paper. In Section 3 I describe two readings that are found in counterfactual reports with counter-identity – what I call the *counterfactual-self* reading, and the *belief-self* reading. I describe three possible accounts of these readings, one of which relies on de se binding. Section 4 presents evidence for a hitherto unobserved correlation between de se/de re construal, and counterfactual-self and belief-self readings – what I call the *De Se Generalization*. I argue that of the three accounts discussed in Section 3, only the one that assumes that there is de se binding can provide a principled explanation of this discovery. Section 5 shows that the resulting picture provides a solution to a longstanding puzzle about so-called *de re blocking effects*, and considers a potential challenge from data involving counter-identity about individuals other than the attitude holder. Section 6 is the conclusion.

2 Theories of de se construal

In this section, I describe in greater detail the two different approaches to de se.

2.1 De se binding

Within the body of literature that posits de se binding in the grammar, we find two broad approaches to how exactly this should be implemented. One approach holds that there are *dedicated de se LFs*, involving binding of the de se pronoun by an abstractor in the left periphery of the complement clause (Chierchia 1990). To illustrate this, I shall assume an extensional framework where worlds are represented in the syntax as unpronounced indexed elements (pronouns) that are bound by co-indexed abstraction operators higher in the structure. Thus the LF for (2) (on its de se reading) is as in (6).

(6) Jonathan believes [$_{\text{CP}}$ λx_1 λw_2 [w_2 he_1 is a loathsome idiot]].

This provides an implementation of Lewis's (1979) idea that to believe de se that you are P is to self-ascribe the property P. If I believe that I am looking at a beautiful view right now, then I self-ascribe the property of looking at a beautiful view. In our story, Jonathan did not self-ascribe the property of being a loathsome idiot; rather he ascribed it to the individual that he was looking at, not realizing that that individual was him.

In (6), binding of the pronoun results in the CP being interpreted as a property rather than a proposition:

(7) $[\![\text{CP}]\!]^{c,g} = \lambda x\ \lambda w$. x is a loathsome idiot in w.

The standard truth conditions of attitude reports require that the embedded clause express a function that when applied to each of the attitudinal alternatives of the attitude holder returns the truth value 1. If the embedded clause in a de se report is a function from individuals to functions from worlds to truth values (type $\langle s, \langle e,t\rangle\rangle$), then attitudinal alternatives cannot be worlds, as on a Hintikka semantics (Hintikka 1969), but rather they are world-individual pairs – so-called 'centred worlds'. The attitudinal alternatives for de se belief, for example (*doxastic* alternatives), can be defined as follows:

(8) $\text{Dox}_{x,w}$ = {$\langle$w', y$\rangle$: it is compatible with what x believes in w for x to be y in w'}.

We correspondingly have the semantics for *believe* given in (9), and the truth conditions for the de se reading of (2) given in (10):

(9) $[\![\text{believe}]\!]^{c,g} = \lambda P\text{: } P \in D_{\langle e,\langle s,t\rangle\rangle}.\ \lambda x\text{: } x \in D_e.\ \lambda w\text{: } w \in D_s.\ \forall\ \langle w',y\rangle \in \text{Dox}_{x,\,w}\text{: } P(y)(w')$

(10) $[\![(2)]\!]^{c,g} = \lambda w.\ \forall\ \langle w',y\rangle \in \text{Dox}_{\text{Jonathan},\,w}\text{: }$ y is a loathsome idiot in w'.

According to (10), for every world-individual pair ⟨w',y⟩ such that for all Jonathan believes, he could be y in w', y is a loathsome idiot in w'. The individual coordinate thus ranges over those inhabitants of Jonathan's belief worlds that he considers to be candidates for himself. Since in our story, Jonathan does not ascribe the property of being a loathsome idiot to any y that is a candidate of Jonathan's for himself, the sentence is correctly predicted to be false on the de se reading.

In languages such as French, German and Italian where *believe* can take an infinitival complement with the unpronounced element PRO in subject position, the resulting report is unambiguously interpreted de se. To account for this, Chierchia proposes that PRO is obligatorily abstracted over:

(11) Jonathan believes $[_{CP}\ \lambda x_1 \lambda w_2\ [w_2\ \text{PRO}_1$ to be a loathsome idiot]].

This is essentially the same LF as (6), and hence receives the same interpretation: PRO is interpreted as a variable ranging over the individual coordinate of Jonathan's doxastic alternatives.

Another way to ensure that PRO is bound by the individual coordinate of the doxastic alternatives of the attitude holder is by positing a lexical entry for PRO that achieves the same effect (Anand & Nevins 2004, Stephenson 2007, 2010). To illustrate this, I will assume now that world variables are represented not in the syntax, but rather as a parameter on the interpretation function. Suppose furthermore, that the evaluation index consists not only of a world parameter, but also of an individual parameter (type e). Evaluation indices are then world-individual pairs, and sentence intensions are sets of such pairs. An attitude verb is a quantifier over world-individual pairs as before, but its first argument is now a sentence intension (hence no need for a lambda abstractor in the syntax), and it binds the world and individual parameters on the interpretation function for the embedded clause:

(12) $[\![.]\!]^{c,g,\langle w,\,x\rangle}$

(13) $[\![\text{u believes S}]\!]^{c,g,\langle w,\,x\rangle} = 1$ *iff* $\forall\ \langle w',y\rangle \in \text{Dox}_{u,\,w}\text{: } [\![S]\!]^{c,g,\langle w',\,y\rangle}$

An attitude verb like *believe* binds the world and individual parameters with respect to which the embedded clause is interpreted. So if some element in the scope of an attitude verb is assigned the individual parameter as its semantic value, then it will be interpreted as a variable ranging over the individual coordinate of the doxastic alternatives of the attitude holder, resulting in de se truth conditions. Anand & Nevins 2004 and Stephenson 2007, 2010 propose that obligatorily controlled PRO is just such an element:

(14) $[\![\text{PRO}]\!]^{c,g,\langle w,\, x\rangle} = x$

(15) $[\![\text{Jonathan believes } [_{CP} \text{ PRO to be a loathsome idiot}]]\!]^{c,g,\langle w,\, x\rangle} = 1$ *iff* $\forall\ \langle w', y\rangle \in \text{Dox}_{\text{Jonathan},w}$: y is a loathsome idiot in w'

The term 'de se binding' as used in this paper refers to a mechanism that results in a pronoun being bound by the individual coordinate of the world-individual pairs ranged over by the attitude verb, where binding is brought about either by lambda abstraction, or as a consequence of the lexical entry of that pronoun.[3]

2.2 De se as a special case of de re

An alternative to de se binding is to ask whether de se construal can be derived via some other mechanism that is independently required.[4] The apparatus involved in the derivation of truth conditions for de re belief reports looks like a good candidate. Remember Quine's old example:

(16) a. Ralph believes that Ortcutt is a spy.

b. Ralph believes that Ortcutt is not a spy.

If we heard these sentences out of the blue and took them to be true, then we would conclude that Ralph is irrational: his belief state seems to ascribe incompatible properties to the same individual. But Quine noticed that in the following scenario, the two sentences are true even though Ralph is not irrational:

3 I set aside the possibility of deriving a de se construal of an ordinary pronoun via a lexical entry that assigns the pronoun the individual parameter as its semantic value; so far as I know such an approach has not been pursued in the literature.

4 In this paper, I focus on the alternative strategy that I will refer to as *de se as a special case of de re*, as it has been pursued in the linguistic literature. However, there is a close connection between this tradition and a philosophical tradition of *de se skepticism* that denies that putative puzzles about de se are anything other than instances of a broader phenomenon of substitution failure in attitudinal environments (Boër & Lycan 1980, Cappelen & Dever 2013, Magidor 2015).

(17) ‘There is a certain man in a brown hat whom Ralph has glimpsed several times under questionable circumstances on which we need not enter here; suffice it to say that Ralph suspects he is a spy. Also there is a grey-haired man, vaguely known to Ralph as rather a pillar of the community, whom Ralph is not aware of having seen except once at the beach. Now Ralph does not know it but the men are one and the same [Bernard Ortcutt].’
(Quine 1956:56)

Intuitively, (16a) is true in this situation because (i) Ralph believes that the person that he saw in the brown hat is a spy and (ii) that person is Ortcutt. (16b) is true because (i) Ralph believes that the person that he saw at the beach is not a spy and (ii) that person is Ortcutt. These facts can hold simultaneously without Ralph being irrational because he does not believe that the person that he saw in the brown hat is identical to the person that he saw at the beach. Roughly speaking, *the person Ralph saw in the brown hat* and *the person Ralph saw at the beach* pick out distinct individuals at each of Ralph’s belief worlds, even though they pick out the same individual in the actual world. Thus Ralph’s belief state ascribes the incompatible properties of being a spy and not being a spy to distinct individuals; perfectly coherent.

According to this account, the truth conditions of the reports of Ralph’s de re belief about Ortcutt are mediated by the kinds of objects that serve as the semantic value of definite descriptions – namely, individual concepts. (For now, we can just think of individual concepts as functions from worlds to individuals, although that assumption will be amended in a moment.) The concepts C that play this mediating role must be *acquaintance-based* in that there must be an acquaintance relation R such that for every world w’ in the domain of the concept, R holds in w’ between the attitude holder and C(w’).[5] Thus the concept expressed by *the person Ralph saw at the beach* is based on the acquaintance relation ‘see at the beach’; that expressed by *the person Ralph saw in the brown hat* is based on the acquaintance relation ‘see in the brown hat’ and so on.

Reinhart’s idea, building on Lewis, was that de se reports are de re reports about the attitude holder that are mediated by a concept that is based on the acquaintance

5 I won’t have much to say about what exactly acquaintance relations are. Let’s just assume that they are relations holding between attitude holders and individuals, that play a causal role in the attitude holder coming to hold a belief about that individual (‘saw at the beach’, ‘heard about from a friend’, ‘read a paper by’, and what have you). Underlying the approach is the assumption that you can only have a de re belief about an individual that you stand in some acquaintance relation to, and that this is reflected in the truth conditions of de re belief reports. Kaplan 1968 also assumed that the acquaintance relations must be sufficiently ‘vivid’, however I will set that aside.

relation SELF. In order to see how this works, we firstly need to flesh out the truth conditions for de re belief reports.[6]

The first task is to determine what the attitudinal alternatives for de re belief are. One might have hoped (as Reinhart seems to have done) that one could revert to treating these as worlds, rather than appealing to the centred worlds discussed in the previous sub-section. But this will not do, for reasons discussed in (Anand 2006). In Quine's scenario, the belief of Ralph's in virtue of which (16a) is true is not 'the man Ralph saw at the beach is a spy', but rather 'the man *I* saw at the beach is a spy'. If Ralph did indeed see Ortcutt at the beach, and in addition he believes (de se) that he is Ronald and that some other guy is Ralph, and he says sincerely *the man Ralph saw at the beach is a spy*, then we are clearly not entitled to conclude that Ralph believes that Ortcutt is a spy. This shows that in Ralph's belief worlds w', it is not Ralph himself who bears the acquaintance relation to the individual returned by the concept in w', but rather those individuals that Ralph is prepared to designate with the first person pronoun. Such individuals are just Ralph's candidates for himself – precisely those individuals designated by the individual coordinate of a centred world.

What is needed is to let the domain of acquaintance-based concepts be world-individual pairs, not worlds.[7] We can then say that if C is an acquaintance-based concept, then there is an acquaintance relation R such that for every world-individual pair $\langle$w', y$\rangle$ that is in the domain of C, R holds between y and C(w', y). If we are to capture the semantics of de re belief reports, then included in the domain of a suitable concept should be the attitude holder's doxastic alternatives – we want a concept that gives us access to the individual that Ralph's *belief* state designates as the guy he saw in the brown hat, for example. So doxastic alternatives – the elements over which *believe* quantifies – must be world-individual pairs, and not merely worlds. We can then state the truth conditions of *Ralph believes that Ortcutt is a spy* as follows.

(18) *Ralph believes that Ortcutt is a spy* is true in w with respect to a contextually supplied concept C with domain $\text{Dox}_{\text{Ralph,w}} \cup \{\langle \text{w, Ralph} \rangle\}$ *iff* C is a suitable

6 It will suffice for the purposes of this paper just to give the truth conditions, without committing ourselves to a particular view on the compositional semantics that derives them. For discussion, see (Charlow & Sharvit 2014, Cresswell & von Stechow 1982, Percus & Sauerland 2003a).

7 An anonymous reviewer suggests that once one concedes that centred worlds are needed to handle de re readings, the case for de se skepticism (see footnote 3) is undermined. It seems to me that in the face of this argument, there are two options available to the upholder of de se skepticism: (i) deny that the theory sketched is the correct theory of de re; or (ii) accept the need for centred worlds in an analysis of de re reports, but maintain that this does not entail that de se readings are in any sense 'special', since on this view the role of centred worlds is not limited to deriving de se readings.

concept of Ortcutt for Ralph and for every $\langle$w', y$\rangle \in \text{Dox}_{\text{Ralph,w}}$: C(w', y) is a spy in w'.

(19) *Suitability*
A concept C is a suitable concept of a res u for an attitude holder x in w *iff*:

(i) C is *acquaintance-based*; and

(ii) C is a *reliable* concept of u for x in w.

(20) *Acquaintance-based*
A concept C is an acquaintance-based concept iff there is an acquaintance relation R such that for every world-individual pair $\langle$w', y$\rangle$ that is in the domain of C, R holds in w' between y and C(w', y).

(21) *Reliability*
A concept C is a reliable concept of u for x in w *iff* C(w, x) = u

Intuitively, the res in a de re report is just the individual that the reported attitude is about – Ortcutt in this case. *Ralph believes that Ortcutt is a spy* is true with respect to C only if in addition to being acquaintance-based, C is a *reliable* concept of Ortcutt. This captures the fact that the truth of *Ralph believes that Ortcutt is a spy* is dependent upon the fact that the man that Ralph saw in the brown hat is Ortcutt: C must return Ortcutt when applied to the pair consisting of the actual world and Ralph. In a world in which that man is in fact Guy (and Ralph still believes that the man he saw in a brown hat is a spy), C is not a *reliable* concept, and hence the sentence turns out to be false (relative to C).

So much for the truth conditions of one particular de re belief report. In order to show how de se can in general be treated as a special case of de re, we need to state the truth conditions for an arbitrary de re belief report ϕ. Let us do that now:

(22) *Truth conditions for de re belief reports*
Let ϕ be a report of a belief held by an attitude holder x that ascribes a property P to a res u. ϕ is true in w with respect to a contextually supplied concept C with domain $\text{Dox}_{\text{x,w}} \cup \{\langle \text{w, x} \rangle\}$ *iff* C is a concept of u that is suitable for x in w and for all $\langle$w', y$\rangle \in \text{Dox}_{\text{x,w}}$: C(w', y) is P in w'.

According to the view that we are considering, a de se belief report is just a de re belief report whose truth is evaluated with respect to a concept C that is based on the acquaintance relation SELF, where SELF is defined as follows.[8]

(23) *SELF*
SELF(x, y, w) *iff* (i) x is sentient in w and (ii) x = y.

We can then define a notion of SELF-based concept as follows.

(24) *SELF-based concept*
C is a *SELF-based concept* iff for every world-individual pair ⟨w', y⟩ that is in the domain of C, SELF holds between y and C(w', y) in w'.

Notice that (24) together with the definition of SELF entails that if C is a SELF-based concept, then for every ⟨w', y⟩ in the domain of C, C(w', y) = y. If C takes as an argument a centred world ⟨w', y⟩ that is a doxastic alternative of the attitude holder, it will return the attitude holder's 'doxastic centre' y. To see how this derives the de se 'construal' of (2), let's assume that the sentence is assigned the following interpretation:

(25) *Jonathan$_i$ believes that he$_i$ is a loathsome idiot* is true in w with respect to a contextually supplied concept C with domain $\text{Dox}_{\text{Jonathan,w}} \cup \{\langle \text{w, Jonathan} \rangle\}$ *iff* C is a concept of Jonathan that is suitable for Jonathan in w and for all ⟨w', y⟩ ∈ $\text{Dox}_{\text{Jonathan,w}}$: C(w', y) is a loathsome idiot in w'.

Given the definition in (24), if the concept with respect to which the sentence is evaluated is SELF-based, then the sentence is true just in case (i) Jonathan is sentient and he is Jonathan and (ii) at each of Jonathan's doxastic alternatives ⟨w', y⟩, y is a loathsome idiot in w'. Setting aside the trivially satisfied condition in (i), this is equivalent to the truth conditions assigned to the dedicated de se LF postulated in Section 2.1.

2.3 Comparison of the two approaches

At least as far as belief reports go, it seems that one can conclude that the de se binding and de se-as-de re routes derive equivalent truth conditions. On the version

8 This is slightly different from the usual definition of SELF, in as much as I have added the condition that SELF(x, y, w) holds only if x is sentient. Typically, SELF is defined simply as identity, but notice that identity is not actually an acquaintance relation: this pen in my hand is identical to itself but it is certainly not *acquainted* with itself. The idea underlying the definition in (23) is that SELF is just that acquaintance relation that all sentient individuals hold towards themselves simply in virtue of being sentient. I thank Frank Sode for discussion of this point.

of the de se-as-de re route presented here, the SELF acquaintance relation that is responsible for de se interpretation is not imposed by the grammar. Rather, the interpretations that the grammar assigns to de re reports are underdetermined with respect to the value of the concept variable C, which is instead contextually supplied. If the context picks a SELF-based concept, then the result will be a sentence that is judged false in a situation where the reported attitude does not satisfy the awareness condition. This approach has the advantage that the only grammatical machinery that it depends on is independently needed for the analysis of de re belief reports; the SELF relation has generally been assumed to come for free: it is simply that acquaintance relation that all sentient individuals bear to themselves.

Until recently, an important argument that de se binding is needed was the existence of anaphoric expressions that are unambiguously read de se (Chierchia 1990). It was reasoned that if de se-as-de re is the only route to de se construal, and if furthermore the SELF acquaintance relation is contextually supplied, then we should not expect the grammar itself to generate attitude reports that can only report de se attitudes. Obligatorily controlled PRO is the classic example of an expression that is unambiguously read de se (Morgan 1970, Chierchia 1990): in languages such as German, French and Italian where *believe* is a control predicate, the counterpart of (26) can only be heard as false in the mirror situation.

(26) John believed PRO to be a loathsome idiot.

Furthermore, logophoric pronouns – dedicated pronominal forms that can only occur in the scope of an attitude verb and obligatorily denote the attitude holder – were taken to be overt counterparts of obligatorily controlled PRO, and thus to provide evidence that some languages have overt expressions that are unambiguously read de se (Heim 2001, 2002, Schlenker 1999, von Stechow 2002, 2003).

Recently, two significant challenges to this view have arisen. Firstly, fieldwork with native speakers has shown that the assumption that logophoric pronouns cannot be read de re is incorrect, at least for the West African language Ewe (Pearson 2012, O'Neill 2016). This suggests at least that the array of unambiguously de se expressions in natural language may be more restricted than was previously thought. Secondly, several authors have developed lines of work where the grammar itself rather than the context imposes the restriction to SELF-based concepts (Schlenker 2003, Maier 2011, Landau 2018, 2015, Santorio 2014). On these approaches, the unavailability of a de re reading for PRO is a truth conditional or presuppositional constraint on the kind of concepts that can mediate the attitude holder's de re belief about herself. If they are correct that the resulting interpretation is truth conditionally equivalent to that arising through de se binding, then the argument for the latter view from obligatory control is undermined.

Furthermore, Landau argues that a de se-as-de re treatment of PRO is to be preferred on the grounds that it permits a return to the traditional approach to control, whereby PRO is bound by the controller. Focusing on Chierchia's approach, he points out that de se binding breaks the traditional dependency between the controller and PRO, which begs the question of how PRO comes to bear the same phi-features as its controller. According to Landau, there is no adequate way of implementing feature transmission from controller to PRO if PRO is bound by an abstractor in the left periphery of the infinitive itself rather than being bound by the controller directly. I take it that Anand & Nevins' and Stephenson's accounts fare no better in this respect since for them too, PRO is not bound by the controller.

In response to the challenge presented by de se-as-de re analyses of PRO (which could potentially also be applied to other unambiguously de se elements such as shifted indexicals), I shall present novel data concerning the interpretation of unambiguously de se expressions in counterfactual attitude reports with counter-identity. The interpretation of these expressions in these environments differs systematically from that of pronouns that are ambiguous between a de se and de re construal. I propose a descriptive generalization ('the De Se Generalization') to capture the facts in this domain, and argue that in order to provide a principled explanation of this discovery, it is necessary to posit de se binding.

3 Counterfactual attitudes with counter-identity

3.1 Introduction

Consider the following pair of sentences.

(27) a. Ivanka imagined that she was Melania and she was giving an interview as First Lady.

b. Ivanka imagined that she was Melania and she was giving an interview as First Daughter.

The sentence *She gave an interview as First Lady* carries the presupposition that the referent of *she* is First Lady. On a natural reading of (27a), this presupposition projects across *imagine*, so that *she* is understood as referring to Melania;[9] the sentence reports Ivanka's imagining herself, as Melania, giving an interview.[10]

9 Actually, the account that will be defended in this paper has as a consequence that the pronoun does not refer at all, but rather is a bound variable. I shall continue to use the terms 'refer' and 'pick out' to indicate the reading that I am interested in; these terms are not intended to signal that the use of the pronoun is referential in any technical sense.

10 I assume that the world is the way that it is at the time of writing: Melania Trump is the United States First Lady, and Ivanka Trump is First Daughter.

Likewise, on a natural reading of (27b) she refers to Ivanka. A scenario in which (27b) might be judged true, for example, is one where Ivanka wonders what it is like for Melania to watch her (Ivanka) giving an interview, and imagines herself as Melania watching such an event. It seems that *she* in these sentences is ambiguous: it refers to either the person who the attitude holder imagines that she is, or the person that she believes that she is (I assume that Ivanka is not mistaken about her own identity.) I will call the former construal the *counterfactual-self* construal, and the latter the *belief-self* construal.

How do the counterfactual-self and the belief-self readings arise? An answer to this question must be compatible with what is known about de re construal in the scope of counterfactual attitudes; in particular, it should be reconcilable with the so-called *puzzle of counterfactual de re* (Ninan 2008) . I describe this puzzle in the next sub-section. Then, I consider three possible views on the source of counterfactual-self and belief-self readings. The first relies on de se binding, while the second and third do not. I shall argue that only the first provides a principled explanation of a hitherto unnoticed empirical generalization, which I call the De Se Generalization:

(28) *The De Se Generalization*

(i) If a pronoun or anaphor is unambiguously read de se, then it cannot receive a belief-self reading in counterfactual reports with counter-identity.

(ii) If a pronoun or anaphor is ambiguous between a de se and a de reading, then it can receive either a counterfactual-self or a belief-self reading in counterfactual reports with counter-identity.

The empirical evidence for the De Se Generalization is set out in detail in Section 4.

3.2 The puzzle of counterfactual de re (Ninan 2008)

We will need a semantics for *imagine*. For now, I will simply give this verb a parallel treatment to that discussed for *believe*: it is a quantifier over elements of a certain set of centred worlds, which in this case I shall refer to as *imagination alternatives*. Roughly speaking, an attitude holder x's imagination alternatives are those world-individual pairs ⟨w', y⟩ such that those things that x imagines to be true are true in w', and y is the inhabitant of w' that x imagines herself to be. Here is the lexical entry.

(29) *De se* imagine (first attempt)
$[\![\text{imagine}]\!]^{c,g,\langle w,\,x\rangle} = \lambda P\text{: } P \in D_{\langle e,\,\langle s,t\rangle\rangle}.\ \lambda x\text{: } x \in D_e.\ \lambda w\text{: } w \in D_s.\ \forall\ \langle w', y\rangle \in \text{Imagine}_{x,\,w}\text{: } P(y)(w')$
Where $\text{Imagine}_{x,w} = \{\langle w', y\rangle :$ it is compatible with what x imagines in w for x to be y in w'$\}$

What about de re imagination? While studies of the semantics of de re reports have typically focused on de re belief, it seems to have been tacitly assumed that the analysis can carry over to other attitude predicates. Suppose we adopt this strategy, and attempt to give schematic truth conditions for de re imagination reports, based on those for de re belief reports given in Section 2.2 above:

(30) *Truth conditions for de re imagination reports* (first attempt)
Let ϕ be a report of an imagining by an attitude holder x that ascribes a property P to a res u. ϕ is true in w with respect to a contextually supplied concept C with domain $\text{Imagine}_{x,w} \cup \{\langle w, x\rangle\}$ *iff* C is a concept of u that is suitable for x in w and for all $\langle w', y\rangle$ in $\text{Imagine}_{x,w}$ C(w', y) is P in w'.

The strategy gives rise to a puzzle, which Ninan describes by inviting us to consider the following sentence.

(31) Ralph is imagining that Ortcutt isn't at the beach.
(Based on Ninan 2008, 2012)

Suppose that there is only one way in which Ralph is acquainted with Ortcutt: he sees him at the beach. It seems that we can conceive of perfectly sensible situations that have this property in which the sentence is true. Indeed, this may be the exact moment at which Ralph sees Ortcutt at the beach; if at that very moment, Ralph imagines that the guy he is looking at is at home watching TV, then we can judge the sentence as true. Yet given (30), the truth conditions for (31) should be as follows.

(32) *Ralph is imagining that Ortcutt isn't at the beach* is true in w with respect to a contextually supplied concept C with domain $\text{Imagine}_{\text{Ralph,w}} \cup \{\langle w, \text{Ralph}\rangle\}$ *iff* C is a concept of Ortcutt that is suitable for Ralph in w and for all $\langle w', y\rangle$ in $\text{Imagine}_{\text{Ralph,w}}$ C(w', y) isn't at the beach in w'.

Since the only acquaintance relation that Ralph bears to Ortcutt is the 'sees at the beach' relation, then if (32) is to stand a chance of being true it must be interpreted with respect to a concept C_{BEACH} that has the following properties:

(33) a. For every world-individual pair ⟨w', y⟩ that is in $\text{Imagine}_{\text{Ralph,w}} \cup \{\langle \text{w, Ralph}\rangle\}$, y sees C_{BEACH}(w', y) at the beach in w'. (Since if C_{BEACH} is a suitable concept of Ortcutt for Ralph, then it must be acquaintance-based.)

b. C_{BEACH}(w, Ralph) = Ortcutt (Since if C_{BEACH} is a suitable concept of Ortcutt for Ralph, then it must be a reliable concept of Ortcutt.)

Furthermore, if (32) is true in the situation that we are considering, then the following must be true.

(34) $\forall \langle \text{w', y}\rangle \in \text{Imagine}_{\text{Ralph,w}}$: C_{BEACH}(w', y) isn't at the beach in w'.

But (33a) and (34) jointly entail that Ralph is imagining that the person that he sees at the beach is not at the beach – that is, that he is imagining something impossible. Yet intuitively, when Ralph sees Ortcutt at the beach and at the same moment imagines that Ortcutt is not at the beach, what he imagines is not impossible.

The lesson is that the truth conditions for de re imagination reports cannot be stated by simply adopting the truth conditions for de re belief reports and replacing doxastic alternatives with imagination alternatives as I did in (30). Instead, it seems that the concepts that mediate the truth conditions of de re imagination reports need to somehow be anchored to the doxastic alternatives of the attitude holder, not his imagination alternatives. Intuitively, *Ralph is imagining that Ortcutt isn't at the beach* is true in our scenario because Ralph is imagining that the man he *believes* that he is looking at the beach isn't at the beach. If we grant that an individual can have the property of being at the beach in Ralph's belief worlds, but lack that property in his imagination worlds, then no contradiction arises.

There are various implementations of this idea on the market (Anand 2011, Ninan 2008, 2012, Yanovich 2011). For concreteness, I shall adopt one of the two solutions considered in Ninan 2008. Ninan's starting point is the assumption that acts of imagining are always relative to belief states, in the sense that what is entertained is counterfactual relative to the doxastic alternatives of the attitude holder. This is modeled by treating the attitudinal alternatives quantified over by *imagine* as pairs of centred worlds ⟨⟨w', y⟩,⟨w'', z⟩⟩, where the first member of the pair, ⟨w', y⟩, is a doxastic alternative of the attitude holder, and the second member of the pair, ⟨w'', z⟩ is compatible with what the attitude holder imagines relative to ⟨w', y⟩.[11] The truth conditions for de re imagination reports can then be stated as follows:

11 I will refrain from examining in detail the question of what it means to imagine something relative to some doxastic alternative. The basic idea is that the interpretation of certain attitudes, such as *imagine*, *wish* and *want* is parasitic on a doxastic modal base (Anand 2011, Heim 1992, Maier 2015, Ninan 2008, Yanovich 2011). This property has been argued to play a role, for example, in the presupposition projection properties of the verbs in question (Heim 1992, Maier 2015).

(35) *Truth conditions for de re imagination reports* (revised version, based on Ninan 2008):[12]
Let ϕ be a report of an imagining by an attitude holder x that ascribes a property P to a res u. ϕ is true in w with respect to a contextually supplied concept C with domain $\text{Dox}_{x,w} \cup \{\langle w, x\rangle\}$ *iff* C is a concept of u that is suitable for x in w and for all $\langle\langle w', y\rangle, \langle w'', z\rangle\rangle$ in $\text{Imagine}_{x,w}$ C(w', y) is P in w''.
Where $\text{Imagine}_{x,w}$= $\{\langle\langle w', y\rangle, \langle w'', z\rangle\rangle$: $\langle w', y\rangle \in \text{Dox}_{x,w}$ and it is compatible with what x imagines in w relative to $\langle w', y\rangle$ for x to be z in w''$\}$
(Based on Ninan 2008: pages 44-45)

The correct truth conditions for (31) are given in (36).

(36) *Ralph is imagining that Ortcutt isn't at the beach* is true in w with respect to a contextually supplied concept C with domain $\text{Dox}_{x,w} \cup \{\langle w, x\rangle\}$ *iff* C is a concept of Ortcutt that is suitable for Ralph in w and for all $\langle\langle w', y\rangle, \langle w'', z\rangle\rangle$ in $\text{Imagine}_{\text{Ralph},w}$ C(w', y) isn't at the beach in w''.

According to the revised truth conditions, the sentence is true with respect to C_{BEACH} only if for all of Ralph's imagination alternatives $\langle\langle w', y\rangle, \langle w'', z\rangle\rangle$, $C_{\text{BEACH}}(w', y)$ isn't at the beach in w'' – that is, if the individual that Ralph believes to be the guy he sees at the beach is not at the beach at those worlds in which what Ralph counterfactually entertains (relative to what he believes) is true. We have now succeeded in stating truth conditions that match our intuitions about the meaning of the sentence. With this background in place, I now proceed to the question of how counterfactual-self and belief-self readings arise.

3.3 First account: De se binding & de se-as-de re (Ninan 2008)

Recall our pair of counterfactual reports:

(37) a. Ivanka imagined that she was Melania and she was giving an interview as First Lady.

b. Ivanka imagined that she was Melania and she was giving an interview as First Daughter.

On the most natural construal of (37a) given world knowledge, the second pronoun refers to Melania (the *counterfactual-self* reading); on the most natural

12 Ninan does not explicitly restrict the domain of the concept as I have done here, in keeping with the formulation of the semantics of de re reports that is adopted in this paper.

construal of (37b), it refers to Ivanka (the *belief-self* reading). Ninan 2008 provides an account of the counterfactual-self and belief-self readings whereby the former arises via de se binding, and the latter arises via de se as a special case of de re construal.[13]

We will need a semantics for de se *imagine*, to maintain the idea that it is a quantifier over pairs of centred worlds:

(38) *De se* imagine (final version)
$[\![\text{Imagine}]\!]^{c,g} = \lambda P\text{: } P \in D_{\langle e, \langle s,t\rangle\rangle}.\ \lambda x\text{: } x \in D_e.\ \lambda w\text{: } w \in D_s.\ \forall\ \langle\langle w', y\rangle, \langle w'', z\rangle\rangle \in \text{Imagine}_{x,\,w}\text{: } P(z)(w'')$

Suppose the de se LF in (40) is assigned to the sentence in (39).[14][15]

(39) Ivanka is imagining (that she is Melania and) she is giving an interview.

(40) Ivanka is imagining $[_{CP}\ \lambda x_1 \lambda w_2\ [w_2\ \text{she}_1\ \text{is giving an interview}]]$

For the purposes of the interpretation of a de se bound pronoun, the first member of the ordered pair of centred worlds quantified over by *imagine* is an idle wheel; the property expressed by the embedded clause is applied to the individual and world coordinates of the second member of this pair:

(41) *De se* imagine (final version)
$[\![(39)]\!]^{c,g} = \lambda w.\ \forall\ \langle\langle w', y\rangle, \langle w'', z\rangle\rangle \in \text{Imagine}_{\text{Ivanka},\,w}\text{: } z \text{ is giving an interview in } w''.$

(39) is thus predicted to be judged true just in case at each of Ivanka's imagination alternatives, the individual that she imagines herself to be is giving an interview. Given the material in parentheses in (39), any y that is the individual coordinate of

13 Ninan's core evidence for the existence of the counterfactual-self/belief-self ambiguity is (i).

(i) I'm imagining that I am Brigitte Bardot and that I am kissing me.
(Ninan 2008: 25, ex 5a; based on Lakoff 1970)

On the intended reading, *I* refers to Brigitte Bardot and *me* refers to the speaker.

14 The material in parentheses in (39) merely establishes the context as one where Ivanka puts herself in Melania's shoes. I do not give the semantics for this portion of the sentence, but a complete representation would have it that the subject of the first conjunct is also de se bound.

15 Since I am for the moment concerned only with overt pronouns, I confine my attention to the version of the de se binding view where the pronoun is bound by a lambda abstractor.

one of Ivanka's imagination alternatives must be Melania – that is, the counterfactual-self. In general, if a pronoun is de se bound in a counterfactual report with counter-identity, it will be construed as picking out the individual that the attitude holder imagines (wishes, pretends, claims etc.) that she is.

To derive the belief-self reading, Ninan assumes that the pronoun is construed de re with respect to a SELF-based concept:

(42) *Ivanka is imagining that she is giving an interview* is true in w with respect to a contextually supplied concept C with domain $\text{Dox}_{\text{Ivanka,w}} \cup \{\langle \text{w, Ivanka} \rangle\}$ *iff* C is a concept of Ivanka that is suitable for Ivanka in w and for all $\langle\langle \text{w'}, \text{y} \rangle, \langle \text{w''}, \text{z} \rangle\rangle$ in $\text{Imagine}_{\text{Ivanka, w}}$ C(w', y) is giving an interview in w".

If we pick that concept C of Ivanka that is suitable for Ivanka and is SELF-based, then the sentence turns out to be true just in case each of Ivanka's imagination alternatives $\langle\langle \text{w'}, \text{y} \rangle, \langle \text{w''}, \text{z} \rangle\rangle$ is such that y is giving an interview in w". This is just to say that Ivanka is imagining that the individual that she believes herself to be (under normal circumstances, Ivanka) is giving an interview.

Ninan's view is essentially that de se binding and de se-as-de re come apart in counterfactual reports with counter-identity: the former gives rise to the counterfactual-self reading, while the latter gives rise to the belief-self reading. This on its own is not an argument that we need de se binding, however. In order to argue for de se binding, we need to show that an account along the lines of Ninan's proposal is to be preferred to an account that does not rely on de se binding, but rather derives the counterfactual-self reading by other means. In the next two sub-sections, I shall consider two such accounts.

3.4 Second account: Anchoring Self-based concepts to counterfactual worlds

Recall that in order to explain why (43) can be true in a situation where Ralph only bears the 'sees at the beach' acquaintance relation to Ortcutt without it following that Ralph is imagining something impossible, Ninan proposed a semantics that assigns the truth conditions in (44) to de re imagination reports.

(43) Ralph is imagining that Ortcutt isn't at the beach.

(44) Let ϕ be a report of an imagining by an attitude holder x that ascribes a property P to a res u. ϕ is true in w with respect to a contextually supplied concept C with domain $\text{Dox}_{\text{x,w}} \cup \{\langle \text{w, x} \rangle\}$ *iff* C is a concept of u that is suitable for x in w and for all $\langle\langle \text{w'}, \text{y} \rangle, \langle \text{w''}, \text{z} \rangle\rangle$ in $\text{Imagine}_{\text{x,w}}$ C(w', y) is P in w".

Where $\text{Imagine}_{x,w}$ = {⟨⟨w', y⟩, ⟨w", z⟩⟩: it is compatible with what x believes in w for x to be y in w', and it is compatible with what x imagines in w relative to ⟨w', y⟩ for x to be z in w"}

But nothing in the data described so far requires that (44) must be the only available reading for (43). What if in addition to the reading on which the concept that mediates the reported de re imagination is anchored to the attitude holder's doxastic alternatives, there is a second reading where that concept is anchored to the centred worlds that are compatible with what he imagines? That is, what if in addition to (44), the grammar also generates the following truth conditions:

(45) Let ϕ be a report of an imagining by an attitude holder x that ascribes a property P to a res u. ϕ is true in w with respect to a contextually supplied concept C with domain $\text{Img}_{x,w} \cup \{\langle w, x\rangle\}$ *iff* C is a concept of u that is suitable for x in w and for all ⟨⟨w', y⟩, ⟨w", z⟩⟩ in $\text{Imagine}_{x,w}$ C(w", z) is P in w".
Where $\text{Img}_{x,w}$ = {⟨w", z⟩: there is some ⟨w', y⟩ such that ⟨⟨w', y⟩, ⟨w", z⟩⟩ ∈ $\text{Imagine}_{x,w}$}

If this were the case, then it would not be necessary to appeal to de se binding in order to generate the counterfactual-self reading. It would be enough to say that the pronouns in these examples are construed de re with respect to a SELF-based concept, on the reading for de re imagination reports given in (45). On this construal, the relevant property is ascribed to that individual that the attitude holder imagines herself to be identical to – Melania Trump in the examples under consideration.

The task, then, is to show that there is no reading of de re imagination reports where the concept is anchored to the attitude holder's imagined worlds rather than her belief worlds. Ninan 2012 offers the following example in order to establish this point:[16]

(46) Scenario: The only way in which Ralph is acquainted with Ortcutt is via the 'sees sneaking around on the waterfront' relation. Ralph imagines that he sees a unique individual sneaking around on the waterfront, and that that individual is dumping a body into the water.
Ralph is imagining that he saw Ortcutt dumping a body into the water.
False
(Ninan 2012: 20-21, ex 21)

We judge (46) false in this scenario, but the semantics above would predict it to be true, as can be checked by examining the predicted truth conditions:

16 I thank an anonymous reviewer for reminding me of this case.

(47) *Ralph is imagining that he saw Ortcutt dumping a body into the water* is true in w with respect to a contextually supplied concept C with domain $\text{Img}_{\text{Ralph,w}} \cup \{\langle \text{w, Ralph}\rangle\}$ *iff* C is a concept of Ortcutt that is suitable for Ralph in w and for all $\langle\langle \text{w', y}\rangle, \langle \text{w'', z}\rangle\rangle$ in $\text{Imagine}_{\text{Ralph,w}}$ z sees C(w'', z) dumping a body into the water in w''.

The sentence is true with respect to C if and only if: (i) C(w, Ralph) = Ortcutt; (ii) Ralph sees C(w, Ralph) sneaking around on the waterfront; (iii) for every $\langle\langle \text{w', y}\rangle, \langle \text{w'', z}\rangle\rangle$ in $\text{Imagine}_{\text{Ralph,w}}$, z sees C(w'', z) sneaking around on the waterfront in w''; (iv) for every $\langle\langle \text{w', y}\rangle, \langle \text{w'', z}\rangle\rangle$ in $\text{Imagine}_{\text{Ralph,w}}$, C(w'', z) is dumping a body into the water in w''. Each of these conditions is met in the scenario above, and yet the sentence is judged false. I take it that this shows that the truth conditions in (45) are not possible truth conditions for de re imagination reports, and hence that the concept that mediates the interpretation of the res must be anchored to the attitude holder's belief worlds, and not to her imagination worlds (or whatever the relevant counterfactual worlds are).

Pranav Anand (p.c.) points out, however, that it may be necessary to accommodate cases where the concept is anchored to the counterfactual worlds after all. Here is a (slightly modified) version of his example:

(48) Scenario: Bill opens the door without looking, to find John on the other side.
John: Hey. Look through the keyhole first. Imagine if I were a robber!

On the relevant reading, John is not inviting Bill to imagine that John has different properties that he actually has (namely, that he is a robber), but rather to imagine that the person he heard ring the doorbell had in fact turned out to be a robber.

I agree that the reading is there, but it does not seem to me to be a genuine de re reading in the sense of a reading that is mediated by an acquaintance-based concept. For one thing, the relevant construal also occurs in the scope of non-attitudinal modals:

(49) John: Hey. Look through the keyhole first. I could have been a robber!

Perhaps, though, *could* is being construed epistemically here, and such a construal permits acquaintance-based concepts to sneak in, so that the sentence means something like 'For all you knew (at the moment when you opened the door), I (the person who you heard ring the doorbell) was a robber'. But we can adapt the scenario so that there is no relevant acquaintance relation available:

(50) Scenario: Bill and Mary are discussing Sue's surprise party in loud voices. They do not realize it, but someone is standing on the other side of the door. Thankfully, it's John, not Sue.

John: Hey. Lower your voices. I could have been Sue!

If *could* is indeed being construed as meaning roughly 'for all you knew', here, then any acquaintaince relation should be anchored to the past knowing time – a time that is in the past, not the present. But at that time, there was no relevant acquaintance relation – Bill and Mary did not know that anyone was standing there. So the pronoun in these cases seems to be interpreted via an individual concept of some sort – 'I (the person at the door) could have been Sue' or 'Imagine if I (the person at the door) had been a robber', but the concept in question is not acquaintance-based.

We find further examples involving modal quantification without any attitude in Nunberg 1993:

(51) Justice O'Connor (a Republican Supreme Court Justice):

a. We might have been liberals.

b. If Democrats had won the last few presidential elections, we might have been liberals. (Nunberg 1993:14, ex 18)

I take it then that (48) is not evidence that concepts are sometimes anchored to counterfactual worlds in counterfactual reports.

Still, one might feel inclined to make a brute force stipulation that there is at least one case where concepts can be anchored to counterfactual worlds in counterfactual reports: namely, when the res denotes the attitude holder and the concept with respect to which the sentence is evaluated is SELF-based. If so, then the grammar would be able to generate truth conditions conforming to the following schema:

(52) Let ϕ be a report of an imagining by an attitude holder x that ascribes a property P to x. ϕ is true in w with respect to the contextually supplied concept C_{SELF} with domain $Img_{x,w} \cup \{\langle w, x\rangle\}$ *iff* C_{SELF} is a SELF-based concept of x that is suitable for x in w and for all $\langle\langle w', y\rangle, \langle w'', z\rangle\rangle$ in $Imagine_{x,w}$ C(w'', z) is P in w''.
Where $Img_{x,w} = \{\langle w'', z\rangle$: there is some $\langle w', y\rangle$ such that $\langle\langle w', y\rangle, \langle w'', z\rangle\rangle \in Imagine_{x,w}\}$

Then the counterfactual-self reading will be generated, without appeal to de se binding, as follows:

(53) *Ivanka is imagining that she is giving an interview* is true in w with respect to the contextually supplied concept C_{SELF} with domain $Img_{x,w} \cup \{\langle w, x\rangle\}$ *iff* C_{SELF} is a concept of Ivanka that is suitable for Ivanka in w and for all $\langle\langle w', y\rangle, \langle w'', z\rangle\rangle$ in $Imagine_{x,w}$ C(w'', z) is giving an interview in w''.

In a context where Ivanka is imagining that she is Melania, the pronoun will be construed as referring to Melania; the belief-self reading will arise by picking the variant of de re *imagine* that anchors the concept to the doxastic alternatives of the attitude holder, as on Ninan's view.

This account is admittedly rather unsatisfying: why should it be possible for concepts to be anchored to counterfactual worlds when the res is the attitude holder and the concept is SELF-based, but not in other cases? Still, it cannot directly be shown that there is no de re variant of imagine that would generate such truth conditions: after all, the truth conditions in question are exactly those that arise on the counterfactual-self reading. Instead, I shall argue in Section 4.3 that such an account fails to give a principled explanation for the De Se Generalization. Before doing so, there is a third approach to counterfactual-self and belief-self readings that I should like to examine.

3.5 Third account: Identification functions

Consider again the semantics for de re imagination reports that we are assuming.

(54) Let ϕ be a report of an imagining by an attitude holder x that ascribes a property P to x. ϕ is true in w with respect to the contextually supplied concept C with domain $\text{Dox}_{x,w} \cup \{\langle w, x\rangle\}$ *iff* C is a concept of u that is suitable for x in w and for all $\langle\langle w', y\rangle, \langle w'', z\rangle\rangle$ in $\text{Imagine}_{x,w}$ C(w', z) is P in w''.

An important property of this semantics that I have been ignoring up until now is that it entails that in the worlds w' in which what the attitude holder imagines is true, some property P holds of an individual that is obtained by applying a concept C to a centred world whose world coordinate is distinct from w'. Moreover, the properties that the individual in question has in the counterfactual worlds will be different from those she has in the belief worlds; for instance, Ortcutt is at the beach at Ralph's doxastic alternatives, but not at his imagination alternatives. In Yanovich's work on the puzzle of counterfactual de re, he argues that this necessitates a method of identifying individuals across worlds – that is, of determining who in the relevant counterfactual worlds corresponds to the individual that the concept supplies at the attitude holder's doxastic alternatives (Yanovich 2011). His solution takes the form of a function g_{id}. Setting aside the technical details of Yanovich's proposal in full, the crucial property of g_{id} for our purposes is that it maps an inhabitant of the attitude holder's doxastic alternatives (that is, an individual obtained by applying the de re concept to the doxastic alternative) to a correspondent of that individual in the relevant counterfactual worlds. (It is of type $\langle e,\langle\langle s,e\rangle,e\rangle\rangle$.) In principle, in a counterfactual report with counteridentity the function might, when applied to

the individual that the attitude holder believes herself to be, return that individual that she counterfactually supposes that she is.[17] This would yield a construal that is equivalent to that obtained by de se binding. To see this, take our First Lady sentence as an example.

(55) Ivanka is imagining (that she is Melania and) she is giving an interview as First Lady.

Incoporating g_{id} into the semantics yields the following truth conditions:

(56) *Ivanka is imagining that she is giving an interview as First Lady* is true in w with respect to a contextually supplied concept C with domain $\text{Dox}_{\text{Ivanka,w}} \cup \{\langle \text{w, Ivanka} \rangle\}$ and an identification function g_{id} *iff* C is a concept of u that is suitable for Ivanka in w and for all $\langle\langle \text{w'}, \text{y}\rangle, \langle \text{w''}, \text{z}\rangle\rangle$ in $\text{Imagine}_{\text{Ivanka,w}}$ $g_{id}(\text{C}(\text{w'}, \text{y}))(\text{w''}, \text{z})$ is giving an interview as First Lady in w".

Since I am continuing to consider strategies for defending the view that de se-as-de re can do anything that de se binding can do, I will focus on the interpretation that results from picking a SELF-based concept as the value of C. Then assuming that Ivanka believes (de se) that she is Ivanka, this is the individual that is supplied as the first argument of g_{id}. If in addition g_{id} has the latitude to map the individual that an attitude holder believes that she is to the individual that she imagines that she is, then in the context under consideration, plugging in Ivanka's imagination alternatives returns Melania. The result is equivalent to that obtained by de se binding.

In Section 4.3 I will argue that this proposal, like the proposal to anchor SELF-based concepts to counterfactual worlds, fails to provide a principled explanation of the De Se Generalization.[18] Before doing so, I shall present the empirical evidence for this generalization.

4 The De Se Generalization

4.1 Part I

We have been considering the following pair of sentences.

17 I thank Igor Yanovich for discussion of this point.

18 Of course, one may reject the assumption that an identification function need figure in a treatment of counterfactual de re in the first place. If so, then one competitor to the view advocated in this paper can be eliminated right away; so much the better for me.

(57) a. Ivanka imagined that she was Melania and she was giving an interview as First Lady.

b. Ivanka imagined that she was Melania and she was giving an interview as First Daughter.

Now consider what happens when we replace the second pronoun with PRO:

(58) a. Ivanka imagined PRO being Melania and PRO giving an interview as First Lady.

b. ?Ivanka imagined PRO being Melania and PRO giving an interview as First Daughter.

(58a) is perfectly acceptable and means roughly the same as (57a). (58b), on the other hand, is a little strange: it seems to presuppose that Melania is First Daughter, contrary to our knowledge about the world. The sentence can be rescued if we assume that Ivanka is imagining that Melania is First Daughter – that is, that what is meant is something like, 'Ivanka imagined that she was Melania and Melania was First Daughter and she – as Melania – was giving an interview as First Daughter.' That such a context is required for (58b) but not for (57b) shows that the overt pronoun has a reading that PRO lacks: PRO has the counterfactual-self reading, but not the belief-self reading.

Now, one might suspect that this contrast between the ordinary pronoun and PRO could have something to do with the peculiar interpretive properties of *imagine* when it takes a gerundive complement: as is well-known, such uses give rise to an experiential reading: you can imagine that you are dead, but you cannot imagine being dead, for instance. But we find the same contrast with other counterfactual attitude verbs. In the examples below, suppose that Ivanka is impersonating Melania (over the phone while talking to a White House aide say), and is feeding the aide false information about Melania (the (a) sentences) or about Ivanka (the intended interpretation of the (b) sentences).

(59) *Pretend*

a. Ivanka pretended that she was Melania and she had just given an interview as First Lady.

b. Ivanka pretended that she was Melania and she had just given an interview as First Daughter.

(60) a. Ivanka pretended PRO to be Melania and PRO to have just given an interview as First Lady.

b. ?Ivanka pretended PRO to be Melania and PRO to have just given an interview as First Daughter.

(61) *Claim*

a. Ivanka claimed that she was Melania and she had just given an interview as First Lady.

b. Ivanka claimed that she was Melania and she had just given an interview as First Daughter.

(62) a. Ivanka claimed PRO to be Melania and PRO to have just given an interview as First Lady.

b. ?Ivanka claimed PRO to be Melania and PRO to have just given an interview as First Daughter.

Now, I should like to claim that the unavailability of the belief-self reading for PRO is related to its inability to be read de re. But to maintain this, I will have to rule out an alternative explanation, pointed out to me by David Adger (p.c.). We assume that PRO must be anaphorically dependent on another argument within the sentence. Additionally, the first PRO in the examples above presumably has to refer to Melania. So if the second occurrence of PRO is dependent on the first one, it should be no surprise that it can only receive the counterfactual-self reading. Let's check this by eliminating the first conjunct from the example.

(63) a. Yesterday, Ivanka called up a White House aide impersonating Melania. During the conversation, she pretended PRO to have just given an interview as First Lady/?First Daughter.

b. Yesterday, Ivanka called up a White House aide impersonating Melania. During the conversation, she claimed PRO to have just given an interview as First Lady/?First Daughter.

The judgments still stand: the unavailability of the belief-self reading is a general property of PRO and not confined to occurrences that are anaphoric on another PRO in the same sentence.

I shall argue that the mechanism responsible for the absence of the belief-self reading is the same mechanism that obliges PRO to be interpreted de se. After

all, ordinary pronouns, which do allow the belief-self reading, also allow the de re reading. But this is not the only difference between PRO and ordinary pronouns. PRO is unpronounced, while ordinary pronouns are overt (in English at least). Furthermore, PRO is highly constrained in where it finds its antecedent: the controller must be either the subject or object of the immediately dominating embedding predicate:

(64) John_i claimed that Mary_j pretended $\text{PRO}_{*i/j}$ to be a genius.

What is needed, then, is to find an expression that is unambiguously de se, is pronounced overtly, and can take a long-distance antecedent with multiple embeddings. The shifted indexical *na* in Korean has all three properties (Park 2014).

In a language where the first person pronoun can undergo indexical shift, it can occur in the scope of an attitude predicate such as *say* and refer to the subject of *say* rather than to the actual speaker. Here is an example from Korean:

(65) Mary-ka nay-ka yengwung-ila-ko malhayssta.
Mary-Nom I-Nom hero-be-C said
'Mary said that {I am, Mary is} a hero.'
(Park 2014:2, ex 6)

A series of tests show that a case like (65) is not an instance of direct speech, but rather of genuine shifting of the pronoun in indirect discourse (Park 2014). Furthermore, the shifted pronoun is unambiguously read de se:

(66) Scenario:

a. John took an exam, and later saw the top 10 scorers with the respective ID numbers. He forgot his own ID number, so didn't know who is who. Pointing to the top score, he remarked "This guy is the smartest!" But it turned out that he was talking about himself.

b. John-i nay-ka ceyil ttokttokhata-ko malhayssta.
John-Nom I-Nom most smart-C said
'John_i said that he_i is the smartest.' False
(Park 2014:9, ex 23)

Finally, the shifted pronoun can take a long-distance antecedent across multiple attitude verbs. This is shown by the ability of *na* to refer to John in readings (c) and (d) of the example below.

(67) Tom: John-i Bill-i na$_1$-eykey Mary-ka na$_2$-lul cohahanta-ko
John-NOM Bill-NOM I-to Mary-NOM I-Acc like-C
malhayssta-ko malhayssta.
said-C said-C
Lit. 'John said that Bill said to me that Mary likes me.'

a. *na*$_1$= Tom *na*$_2$= Tom

b. *na*$_1$= Tom *na*$_2$= Bill

c. *na*$_1$= John *na*$_2$= John

d. *na*$_1$= John *na*$_2$= Bill (Park 2014 :17, ex 48)

The question, then, is whether shifted *na* can receive the belief-self reading in a counterfactual report with counter-identity. It cannot, as shown by the falsity of (68) in the following scenario.[19][20]

(68) Scenario:
Mary wants a bit of good publicity for herself, and thinks that an endorsement from Hilary Clinton is the way to get it. Yesterday, she called up a journalist impersonating Hilary Clinton, and during the conversation, said 'Mary is a hero'.
Mary-ka nay-ka yengwung-ila-ko malhayssta.
Mary-NOM I-NOM hero-be-C said
'Mary said that {I am, Mary is} a hero. False

By contrast, the counterfactual-self reading is available: (69) is judged true in the scenario below.

(69) Scenario:
Mary wants to generate a bit of bad publicity for Hilary Clinton, by getting a journalist to write a story about what a show off she is. Yesterday, she called up a journalist impersonating Hilary Clinton, and during the conversation, said 'I am a hero'.

Note that this pattern is specific to the shifted indexical: replacing the first person pronoun with a third person pronoun produces a sentence that is judged true in both scenarios:

19 I thank Yangsook Park and Dorothy Ahn for their judgments about these cases.

20 As is typical for shifting languages, the range of embedding predicates that license indexical shift in Korean is highly limited; the examples discussed here all involve *say*, for the reason that it is the only counterfactual attitude verb that licenses indexical shift.

(70) Mary-ka kunye-ka yengwung-ila-ko malhayssta
Mary-NOM she-NOM hero-be-C said
'Mary said that she is a hero.'

So the belief-self reading is unavailable for both Korean shifted indexicals and obligatorily controlled PRO in counterfactual reports with counter-identity. What these two expressions have in common is that both are unambiguously read de se. I conclude from this that if an expression is unambiguously read de se, it cannot receive a belief-self reading. This is the first piece of the De Se Generalization:

(71) *The De Se Generalization: Part I*
If a pronoun or anaphor is unambiguously read de se, then it cannot receive a belief-self reading in counterfactual reports with counter-identity.

As for expressions that admit both de se and de re readings, we have already caught a glimpse of how they behave: we have seen that ordinary pronouns in both English and Korean allow both counterfactual-self and belief-self readings. In the next sub-section, I will provide further evidence that expressions showing de se/de re ambiguity also display the counterfactual-self/belief-self ambiguity.

4.2 Part II

Reflexives are ambiguous between a de se and a de re reading, even when they are bound by a de se subject (Sharvit 2010):

(72) Jonathan$_i$ said that he$_i$ disliked himself$_i$.
Scenario 1: Jonathan said 'I dislike myself'. True
Scenario 2: Jonathan pointed at his reflection, not realizing that he was looking in a mirror, and said, 'I dislike that guy'. True

Likewise, reflexives display both the counterfactual-self and the belief-self reading (Heim 1994):

(73) a. Ivanka imagined that she was Melania and she was watching herself on TV giving an interview as First Lady.

b. Ivanka imagined that she was Melania and she was watching herself on TV giving an interview as First Daughter.

Secondly, the logophoric pronoun *yè* in Ewe can be read de re as well as de se (Pearson 2015, 2012, O'Neill 2016):[21]

21 A logophoric pronoun is a pronoun that can only occur in the scope of an attitude predicate and must refer to the bearer of the attitude.

(74) John be yè le cleva
John say LOG COP clever
'John said that he was clever.'

Scenario: John has just found an old paper that he wrote, but he doesn't realize that he is the author of the paper. He reads it and is impressed by what a good paper it is. He says, "Whoever wrote this paper is clever". *True*.
(Pearson 2015:98, ex 51)

It too can receive either the belief-self or the counterfactual-self reading:[22]

(75) a. John koudrin be yè nyi Barack Obama koudo yè na
John dream COMPL LOG COP Barack Obama CONJ LOG give
yè cadeau
LOG gift
'John dreamed that he was Barack Obama and he gave himself a gift.'
Scenario: Last night, John dreamt that he was Barack Obama. In the dream, he, as Barack Obama, gave himself (John) a gift. *True*.
(Pearson 2015:101, ex 60)

So if a pronoun or anaphor is ambiguous between a de se and a de re reading, then it is also ambiguous between the counterfactual-self and the belief-self reading. The De Se Generalization can now be stated in its entirety as follows:

(76) *The De Se Generalization*

(i) If a pronoun or anaphor is unambiguously read de se, then it cannot receive a belief-self reading in counterfactual reports with counter-identity.

(ii) If a pronoun or anaphor is ambiguous between a de se and a de reading, then it can receive either a counterfactual-self or a belief-self reading in counterfactual reports with counter-identity.

The De Se Generalization reveals that there is an intimate relationship between the availability of de se/de re readings, and that of counterfactual-self and belief-self readings. In the next sub-section, I shall argue that only an account of counterfactual-self and belief-self readings that appeals to de se binding can provide a principled explanation of this relationship.

22 I assume that *dream* is a counterfactual attitude verb, meaning roughly 'imagine while asleep'. For arguments to this effect, see section 5.1.

4.3 Explaining the De Se Generalization

The De Se Generalization reveals a correlation between the unavailability of de re readings and the unavailability of belief-self readings. As the above discussion shows, this correlation is attested cross-linguistically across a variety of species of pronouns and anaphora. A principled explanation of this discovery will (i) derive both de se and counterfactual-self readings with a single mechanism and (ii) derive both de re and belief-self readings with a single mechanism. A system that does not meet these desiderata will treat the correlation highlighted by the De Se Generalization as a mere coincidence. In this sub-section, I will examine how the three approaches to counterfactual- and belief-self readings discussed in Section 3 fare with respect to desiderata (i) and (ii).

Recall Ninan's (2008) proposal that counterfactual-self readings arise via de se binding, and belief-self readings arise as de re readings mediated by a SELF-based concept. If we assume that unambiguously de se expressions such as PRO and shifted indexicals undergo de se binding (either by a lambda abstractor, or as a consequence of the semantics of the expression itself), then it will automatically follow that such expressions receive counterfactual-self readings but not belief-self readings. An expression that is ambiguous between a de se and a de re reading on the other hand can either be de se bound, or be read de re. The availability of the former mechanism guarantees the availability of the counterfactual-self reading, while the availability of the latter one ensures that the expression can be assigned a belief-self reading. That pronouns and anaphora should conform to the De Se Generalization is exactly what is expected on this approach, with no need for any further assumptions.

Matters are less straightforward, however, if we turn our attention to approaches that derive de se readings as a special case of de re. The first such approach that we discussed accounted for counterfactual-self readings by assuming that these are mediated by a SELF-based concept that is anchored to the counterfactual alternatives of the attitude holder, rather than to her doxastic alternatives. Belief-self readings would then arise where the mediating concept is anchored to the attitude holder's doxastic alternatives. In Section 3.4 I offered reasons for skepticism about the notion that de re concepts can be anchored to the attitude holder's counterfactual alternatives. Let's assume for the moment, though, that this is possible, at least when the res is the attitude holder and the concept in question is SELF-based.

In order to ensure that unambiguously de se expressions cannot be assigned the belief-self reading, proponents of this approach would have to require that the SELF-based concept can *only* be anchored to the counterfactual alternatives of the attitude holder, and not to her doxastic alternatives. But what would ensure that this is the case? There does not seem to be any reason why a de se-as-de re analysis of de se readings should require anchoring of the concept to the attitude holder's

counterfactual alternatives when the embedding predicate is counterfactual; indeed with doxastic predicates, the concept is anchored to the belief worlds of the attitude holder – why shouldn't the same hold for counterfactual predicates as well?

To exclude this possibility, one would have to stipulate that unambiguously de se expressions such as PRO and shifted indexicals enter the semantic composition not only with a requirement that they be construed with respect to a SELF-based concept, but with an additional requirement that when embedded below a counterfactual attitude, this concept is anchored to the counterfactual alternatives of the attitude holder. As Pranav Anand (p.c.) has pointed out to me, such a requirement could perhaps be explained in the case of PRO by appeal to the independent condition that PRO find its antecedent in the immediately dominating clause; the anchoring of the concept to counterfactual alternatives could then be considered a reflex of this locality condition in the domain of (centred) worlds (say, by assuming that counterfactual attitudes decompose into their counterfactual and doxastic components in the syntax). But such an account cannot be applied to shifted indexicals, which we have seen can take a long distance antecedent. No, what PRO and shifted indexicals have in common is that they are both unambiguously read de se. An account of this property that derives the counterfactual-self construal for free should be preferred to one that does not. Obligatory anchoring of a SELF-based concept to counterfactual centred worlds belongs in the latter category.

A de se-as-de re approach appealing to an identification function for mapping individuals at doxastic alternatives to individuals at counterfactual alternatives fares no better: such an approach must assume that the mechanism that ensures that expressions such as PRO and shifted indexicals are construed de re with respect to a SELF-based concept is augmented with a function that maps the belief-self to the counterfactual-self in counter-identity reports. This fails to meet the desideratum of deriving de se and counterfactual-self readings with a single mechanism.

In general, a view that treats de se readings as a special case of de re readings will struggle to explain the De Se Generalization. The crux of the problem is that on such a view, both the counterfactual-self and the belief-self reading are derived by the same mechanism that gives rise to de se readings: de re construal mediated by a SELF-based concept. But in order to meet our desiderata for a principled explanation of the De Se Generalization, we should posit that the mechanism that yields de se readings also produces counterfactual-self readings, with belief-self readings coming about some other way. I conclude that a principled explanation of the De Se Generalization necessitates positing de se binding in the grammar.

5 Consequences of the proposal

In this section I describe two consequences of the view proposed in this paper.

5.1 De re blocking effects

The way of thinking about de se construal in counterfactual reports developed here sheds new light on so-called *de re blocking effects*. Percus & Sauerland 2003b noticed that in dream reports with counter-identity, a de se pronoun cannot be c-commanded by a de re pronoun:

(77) Ivanka dreamed that she was Melania and she disliked her husband.

According to Percus & Sauerland, this sentence can report a dream in which Ivanka's dream self – Melania – dislikes Ivanka's husband. On this reading, *she* is construed de se and *her* de re. However, it cannot report a dream where Ivanka dislikes Donald Trump. The judgements are subtle, but have received initial experimental support from Pearson & Dery 2014.[23]

Several accounts of de re blocking effects have been offered in the literature (Anand 2006, Charlow 2010, Percus & Sauerland 2003b). All of them share the assumption that de re blocking is a form of syntactic intervention, which is expected if de se construal comes about via binding, but is unexplained otherwise. For example, Percus & Sauerland propose that a de se pronoun is not bound in situ, but rather behaves like a relative pronoun in that it moves (covertly) to the left edge of the embedded clause, triggering insertion of an abstractor that binds the trace. The resulting structure is assigned a property interpretation; in this sense, the proposal is a variant of the binding in situ approach:[24]

(78) a. $[_{CP}$ (she) λx_1 λw_2 [w_2 t_1 disliked her$_3$ husband]].

b. $[\![CP]\!]^{c,g} = \lambda x \lambda w$. x dislikes g(3)'s husband in w

A principle of Superiority dictates that given two pronouns that can both undergo this movement operation, it is the structurally higher one that should move. The reason why the sentence cannot be understood as reporting a dream where Sophia dislikes Barack Obama is that this would require movement of the possessive pronoun across the de re subject, which would violate Superiority.

To the extent that this proposal is successful it too provides an argument for de se binding: if there are no syntactic dependencies that can produce de se readings then it is difficult to explain why the array of de se and de re construals is apparently

23 The empirical picture is further complicated by a follow-up study to Pearson & Dery 2014 (Dery & Pearson 2015). This paper shows that participants' judgments in this domain are sensitive to both effects of exposure and task effects, which need to be better understood if the experimental data are to be used as a basis for robust conclusions about linguistically naïve speakers' judgments.

24 I gloss over the details of the de re interpretation of the possessive pronoun, to keep the representations simple.

structurally conditioned. However, a sticking point for this approach is that de re blocking effects apparently only show up with certain predicates. Anand 2006 noticed that in the scope of *believe*, a de re pronoun can c-command a de se pronoun:[25]

(79) Ivanka believed that she was Melania and she disliked her husband.

Suppose that Ivanka is suffering from amnesia, and confused about her identity. Various circumstances lead her to form the belief that she is Melania, the wife of the President. She comes across a newspaper article that she herself wrote, which is very critical of the President. She concludes, 'Wow, whoever wrote this must really dislike my husband'. According to Anand, a sentence like (79) is true in this type of situation. He concludes that de re blocking effects are found with *dream* but not with *believe*.

Since we should not expect the syntactic constraint which gives rise to de re blocking to be active in the scope of some verbs but not others, Anand takes his data to show that there is a second route to de se, in addition to de se binding. This route – de se as a special case of de re – is available for the possessive pronoun in the scope of *believe* but not for *dream*. Consequently, a de re pronoun can c-command a corresponding de se pronoun in a belief report without violation of syntactic constraints.

In this paper I align my support with Anand's conclusion – both routes to de se are needed. In addition, we are now in a position to understand why the de se-as-de re route is apparently available for *believe* but not for *dream*. Anand answers this question by giving a semantics for *dream* that constrains the concepts that may mediate the interpretation of de re material in its scope: such concepts may not be SELF-based. This solution is obviously unsatisfying: no principled explanation is given of why de re construal should be constrained in this way in the scope of *dream* but not of *believe*. [26]

By contrast, the view proposed in this paper has the consequence that nothing special needs to be said about the semantics of *dream* and *believe* in order to explain why the former shows de re blocking effects but the latter does not. I shall assume that the Superiority-based explanation of de re blocking effects proposed by Percus & Sauerland for *dream* is correct. The crucial difference between the two verbs is that whereas in the scope of doxastic predicates with counter-identity, the two routes to de se return the same individual, in the scope of *dream* they do not. As has now been argued at length, de se binding returns the counterfactual self, whereas

25 Again, see Pearson & Dery 2014 for initial experimental support for these judgments.

26 Percus 2006 suggests that the difference between *dream* and *believe* may be that the former but not the latter involves 'putting oneself in someone else's shoes' – precisely the type of construal that I am concerned with in this paper. The view developed in this paper provides an articulation of why this should matter for de re blocking.

de se-as-de re returns the belief self. Consequently, the de se-as-de re route lets the possessive pick out Melania in (79), but not in (77).

Notice that this solution depends on the assumption that *dream* is like *imagine*, *pretend*, *wish* and *claim* in denoting an attitude that an attitude holder can coherently hold towards a content that she believes to be false. This is at odds with the treatment of *dream* given in Percus & Sauerland 2003b, for example, where it is analyzed as meaning roughly 'believe while asleep'. Instead, I propose that the verb should more properly be treated as 'imagine while asleep'. In fact, there is a philosophical literature arguing in favor of this latter view of the nature of dreams (Ichikawa 2009). As linguistic evidence, consider the sentence in (80).

(80) Ralph is dreaming that Ortcutt never existed.

We can re-construct the puzzle of counterfactual de re with this case: (80) can report a de re dream that Ralph is having about Ortcutt (qua the man he saw at the beach, say) that does not involve an impossibility (that the man Ralph saw never existed). So we can run the arguments that we have developed for *imagine* on *dream*: in a report of a dream that Ivanka has in which she is Melania, de se via binding gives back Melania, and de se-as-de re gives back Ivanka. This provides a more principled explanation of how de re blocking effects are conditioned by choice of embedding predicate than was available before now.

5.2 Counter-identity without de se?

In this sub-section, I consider the possibility that the counter-identity cases I have discussed are not specific to de se reports (Dever 2014).

I have argued that in a counterfactual report with counter-identity, the only way for a nominal expression that takes the attitude holder as its antecedent to be construed as picking out the counterfactual self is if it is de se bound. As is well known, a nice property of de se binding is that in counter-identity reports such as those in (81), it circumvents the problem of identity between distinct individuals that would arise on a propositional view: it does not matter that there is no world in which Ivanka is Melania, or where Heimson is Hume, since the embedded pronouns are not interpreted as denoting those individuals, but rather are merely lambda abstracted variables.

(81) a. Ivanka is imagining that she is Melania.

b. Heimson believes that he is Hume.

But de se binding is not the only way of circumventing this problem:

(82) Bill believes that Ivanka is Melania.

Here, the trick is that *Ivanka* and *Melania* are both construed de re; what is required of Bill's doxastic alternatives is not that Sophia and Michelle Obama be identical, but rather that the individuals that the relevant concepts return are. We should ask, then, what happens in cases where *believe* is replaced with a counterfactual attitude:

(83) a. Bill is imagining that Ivanka is Melania.

b. Bill wishes that Ivanka were Melania.

Notice firstly that it is difficult to know what to make of these sentences on encountering them out of the blue. But with a bit of context, they can be heard as acceptable. Suppose, for example, that Bill used to work for Melania, and he now works for Ivanka. He preferred Melania as a boss, so he imagines that he is working for her again, or wishes that he were working for her again. In these scenarios, I think that the sentences in (83) can be judged true.

This is initially surprising given what I have said about counterfactual de re: if Bill is not mistaken about who Ivanka and Melania are, then the sentences should require that in the relevant counterfactual worlds Ivanka is Melania. Is this a problem for our overall picture?

I think that these cases need to be understood better, but that they do not undermine the view developed in this paper. Firstly, it is telling that they require a significant amount of contextual support: out of the blue, I do not know what it means to imagine that Ivanka is Melania, but I do know what it means to imagine that I am Melania. In the former case, there must be some contextually salient role that Ivanka occupies (such as being Bill's boss); in the latter, this is not needed. In fact, this "role-occupying" type of identity predication can arise without an attitude in sight. If I temporarily fill in for Sophia at her job (because she is sick, say), I might say to the staff:

(84) For today, I am Sophia.

Thus one way of dealing with the cases in (83) that leaves our view in tact is to invoke something like the following lexical entry for the copula:

(85) $[\![\text{be}]\!]^{c,g} = \lambda x \lambda y \lambda w$. in w, x occupies the role canonically occupied by y

The view that counterfactual de se with counter-identity is a distinct phenomenon from counter-identity about other individuals is supported by failure of inference from the first type of attitudinal content to the second type. Consider the following example:

(86) Ivanka's best friend has the same desires as Ivanka has.
Ivanka wishes that she were Melania.
#Therefore, Ivanka's best friend wishes that Ivanka were Melania.

This is quite different from the inference patterns that we find in counterfactual reports without counter-identity: in principle, there is nothing that blocks an inference from x desiring (de se) property P to y desiring that P hold of x:[27]

(87) Ivanka's best friend has the same desires as Ivanka has.
Ivanka wishes that she were popular.
Therefore, Ivanka's best friend wishes that Ivanka were popular.

In sum, the data in this section support the view that there is something distinctive about counterfactual de se attitudes with counter-identity: they cannot be collapsed together with reports of imagined or desired counter-identity about third parties. This is to be expected if the former class of case is derived by a binding mechanism that is unavailable for the latter class.[28]

6 Conclusion

Unambiguously de se expressions such as PRO and shifted indexicals have played a starring role in recent debates about the proper analysis of de se construal. At the outset of this paper, it seemed that this emphasis may have been misplaced: successive authors have shown that lack of a de re reading can be accounted for on a de se-as-de re view, by building the requirement that the mediating concept be SELF-based into the semantics. Yet the upshot of the argument developed here is that pronouns and anaphors that can only be read de se *do* have something important

27 An anonymous reviewer reports that they can get a strict reading for an ellipsis case such as (i), whereby it entails that Ivanka's best friend wishes Ivanka were Melania.

(i) Ivanka wishes that she were Melania and Ivanka's best friend does too.

To the extent that the strict reading is available for me, it is dispreferred relative to the sloppy reading. Compare this to (ii), where the strict and sloppy readings seem to be equally easy to access:

(ii) Ivanka wishes that she looked like Melania and Ivanka's best friend does too.

I leave it to future work to investigate these subtle judgments in greater depth.

28 An anonymous reviewer suggests that it may be possible to handle the cases discussed in this subsection by a mechanism similar to that posited in Nunberg 1993 for examples like *The ham sandwich is sitting at table 7* or *I am parked out back*. (See also example (84) above.) I leave it to future work to pursue this possibility further.

to teach us about the routes to de se, in virtue of the correlation between the lack of a de re reading and the lack of a belief-self reading. A principled account of this correlation should appeal to de se binding.

What flavour of de se binding is needed? It seems more plausible to treat ordinary pronouns assigned the counterfactual-self reading as bound by a lambda abstractor in the syntax, rather than this reading arising as a result of the lexical entry of (a particular use of) the pronoun.

After all, we must in any case say that ordinary pronouns can be lambda-bound in order to account for bound-variable uses as diagnosed by sloppy readings under VP-ellipsis, binding by quantifiers, etc. Moreover, this assumption appears to be necessary in order to account for the de se blocking effects observed by Percus & Sauerland and discussed in Section 5.1.

I am inclined to say that PRO too is (mandatorily) lambda-bound. As Chierchia 1990 noted, not all control predicates are attitude verbs; some, such as *force*, are plain modal quantifiers. This poses no problem for a uniform treatment of PRO as a lambda-bound variable: the control infinitive expresses a property which can serve as an argument for either a quantifier over centred worlds (an attitude verb), or a quantifier over worlds (a modal), provided that a suitable lexical entry for the predicate itself is given. By contrast, an Anand & Nevins- or Stephenson-style approach which implements the de se binding of PRO by assigning it a particular lexical entry must say something else entirely about the interpretation of PRO in the scope of a verb like *force*; since such predicates are not quantifiers over centred worlds, the individual parameter on which PRO depends for its semantic value on this view would fail to be bound in such cases.

I conclude, then, that not only is there de se binding, but there are *dedicated de se LFs* – syntactic structures that are unambiguously assigned the truth conditions of a de se report.[29] However, I shall leave as open questions (i) the issue raised by Landau of how PRO is assigned phi-features if it is not bound directly by the controller, and (ii) how de se binding of shifted indexicals is implemented.

If the line of argument developed in this paper is correct, then the debate between proponents of de se binding and of de se as a special case of de re is premised on a mistaken assumption: namely, that the two routes yield equivalent truth conditions in all cases. I have argued that this assumption is incorrect, building on Ninan's insight that the two routes can produce different construals in counterfactual reports with counter-identity. A principled explanation of the De Se Generalization rests on the assumption that the two routes not only can yield distinct interpretations in this class of case, but in fact they must. There is no way for a pronoun that is construed de se-as-de re to be interpreted as picking out the attitude holder's counterfactual

29 Additional arguments for this view can be found in Percus & Sauerland 2003a (the so-called 'argument from *only*; see Anand 2006 for a critique), and Patel-Grosz 2015.

self, for principled reasons having to do with the workings of de re in counterfactual attitude reports. Where we find a pronoun that can be construed in this way, we should conclude that it can be de se bound. Where we find one that can *only* be construed in this way, we should conclude that it must be. Overt pronouns fall into the former category, and PRO and shifted indexicals fall into the latter one.

This line of argument lets us be more precise about the facts that need to be derived by a semantics for sentences that unambiguously report de se attitudes – whether the culprit be PRO, a shifted indexical, or some other pronoun or anaphor. For instance, the core semantic fact about obligatorily controlled PRO is not simply that it is unambiguously read de se, but that in addition it can only receive the counterfactual-self reading in counterfactual reports with counter-identity. In this respect, theories of control that appeal to de se binding rather than de se-as-de re enjoy a clear advantage.

References

Anand, Pranav. 2006. *De De Se*. Massachusetts Institute of Technology dissertation. http://hdl.handle.net/1721.1/37418.

Anand, Pranav. 2011. Suppositional projects and subjectivity. *Ms., University of California at Santa Cruz.* http://web.eecs.umich.edu/~rthomaso/lpw11/anand.pdf.

Anand, Pranav & Andrew Nevins. 2004. Shifty operators in changing contexts. *Semantics and Linguistic Theory* 14. 20–37. https://doi.org/10.3765/salt.v14i0.2913.

Boër, Sam E & William G Lycan. 1980. Who, me? *The Philosophical Review* 89(3). 427–466. http://www.jstor.org/stable/2184397.

Cappelen, Herman & Josh Dever. 2013. *The inessential indexical: On the philosophical insignificance of perspective and the first person*. OUP Oxford. https://doi.org/10.1093/acprof:oso/9780199686742.001.0001.

Charlow, Simon. 2010. Two kinds of de re blocking. *Handout of talk given at MIT Ling Lunch.* https://simoncharlow.com/handouts/mit2up.pdf.

Charlow, Simon & Yael Sharvit. 2014. Bound 'de re' pronouns and the LFs of attitude reports. *Semantics and Pragmatics* 7(3). 1–43. https://doi.org/10.3765/sp.7.3.

Chierchia, Gennaro. 1990. Anaphora and attitudes de se. In J. van Benthem & van Emde Boas In R. Bartsch (ed.), *Semantics and contextual expression*, 1–32. Dordrecht: Foris.

Cresswell, Maxwell J. & Arnim von Stechow. 1982. De re belief generalized. *Linguistics and Philosophy* 5(4). 503–535. https://doi.org/10.1007/BF00355585.

Dery, Jeruen E & Hazel Pearson. 2015. *On experience-driven semantic judgments: A case study on the Oneiric Reference Constraint*. Universitätsbibliothek Tübingen.

https://doi.org/10.15496/publikation-8629.
Dever, Josh. 2014. Reply to Richards' and Pearson's comments on *The Inessential Indexical: on the Philosophical Insignificance of Perspective and the First Person*. October 2014.
Heim, Irene. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of Semantics* 9(3). 183–221. https://doi.org/10.1093/jos/9.3.183.
Heim, Irene. 1994. Puzzling reflexive pronouns in de se reports, http://semanticsarchive.net/Archive/mVkNjZlY/Bielefeld%20handout%2094.pdf.
Heim, Irene. 2001. Semantics and morphology of person and logophoricity. *Handout of talk given at the University of Tübingen* .
Heim, Irene. 2002. Features of pronouns in semantics and morphology. *Handout of a talk given at USC* .
Hintikka, Jaakko. 1969. Semantics for propositional attitudes. In *Models for modalities*, 87–111. Springer. https://doi.org/10.1007/978-94-010-1711-4_6.
Ichikawa, Jonathan. 2009. Dreaming and imagination. *Mind & Language* 24(1). 103–121. https://doi.org/10.1111/j.1468-0017.2008.01355.x.
Kaplan, David. 1968. Quantifying in. *Synthese* 19(1). 178–214. https://doi.org/10.1007/BF00568057.
Lakoff, George. 1970. Linguistics and natural logic. *Synthese* 22(1). 151–271. https://doi.org/10.1007/BF00413602.
Landau, Idan. 2015. *A two-tiered theory of control*, vol. 71. MIT Press. https://doi.org/10.7551/mitpress/9780262028851.001.0001.
Landau, Idan. 2018. *Direct variable binding and agreement in obligatory control* 1–41. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-56706-8_1.
Lewis, David. 1979. Attitudes de dicto and de se. *The Philosophical Review* 88(4). 513–543. https://doi.org/10.2307/2184843.
Magidor, Ofra. 2015. The myth of the de se. *Philosophical Perspectives* 29(1). 249–283. https://doi.org/10.1111/phpe.12065.
Maier, Emar. 2006. *Belief in context: Towards a unified semantics of de re and de se attitude reports*. Radboud Universiteit Nijmegen dissertation. https://sites.google.com/site/emarmaier/publications.
Maier, Emar. 2009. Presupposing acquaintance: a unified semantics for de dicto, de re and de se belief reports. *Linguistics and Philosophy* 32(5). 429–474. https://doi.org/10.1007/s10988-010-9065-2.
Maier, Emar. 2011. On the roads to de se. In *Semantics and Linguistic Theory*, vol. 21, 393–412. https://journals.linguisticsociety.org/proceedings/index.php/SALT/article/viewFile/2611/2358.
Maier, Emar. 2015. Parasitic attitudes. *Linguistics and Philosophy* 38(3). 205–236. https://doi.org/10.1007/s10988-015-9174-z.

Morgan, Jerry. 1970. On the criterion of identity for noun phrase deletion. In *Sixth regional meeting, Chicago Linguistic Society*, 380–389.

Ninan, Dilip. 2008. *Imagination, content, and the self*. Massachusetts Institute of Technology. http://hdl.handle.net/1721.1/45621.

Ninan, Dilip. 2012. Counterfactual attitudes and multi-centered worlds. *Semantics and Pragmatics* 5(5). 1–57. https://doi.org/10.3765/sp.5.5.

Nunberg, Geoffrey. 1993. Indexicality and deixis. *Linguistics and Philosophy* 16(1). 1–43. https://doi.org/10.1007/BF00984721.

O'Neill, Teresa. 2016. The distribution of the Danyi Ewe logophor *yi*. Talk given at the Annual Meeting of the Linguistic Society of America. https://static1.squarespace.com/static/55e3a4d0e4b06cce652896c2/t/56a2ba800e4c11785794d0bd/1453505153503/LSA_O%27Neill_slides.pdf.

Park, Yangsook. 2014. Indexical shift and the long-distance reflexive Caki in Korean. Unpublished Ms. http://blogs.umass.edu/yangsook/files/2011/09/Park_2014_indexical_shifting.pdf.

Patel-Grosz, Pritty. 2015. Pronominal typology and the *de se/de re* distinction. Unpublished manuscript, University of Oslo. http://semanticsarchive.net/Archive/ThlYTM5N/patelgrosz_pro_dere_dese.pdf.

Pearson, Hazel. 2012. *The sense of self: Topics in the semantics of de se expressions*. Harvard University dissertation. http://semanticsarchive.net/Archive/WNhNjVjO/.

Pearson, Hazel. 2015. The interpretation of the logophoric pronoun in Ewe. *Natural Language Semantics* 23(2). 77–118. https://doi.org/10.1007/s11050-015-9112-1.

Pearson, Hazel & Jeruen Dery. 2014. Dreaming de re and de se: Experimental evidence for the Oneiric Reference Constraint. In *Proceedings of Sinn und Bedeutung*, vol. 18, 322–339.

Percus, Orin. 2006. Uninterpreted pronouns? Unpublished manuscript.

Percus, Orin & Uli Sauerland. 2003a. On the LFs of attitude reports. In *Proceedings of Sinn und Bedeutung 7*, University of Konstanz, Konstanz. http://semanticsarchive.net/Archive/jI4NmJlY/PercusSauerland03a.pdf.

Percus, Orin & Uli Sauerland. 2003b. Pronoun movement in dream reports. In *Proceedings of NELS*, vol. 33, 265–284.

Quine, Willard V. 1956. Quantifiers and propositional attitudes. *the Journal of Philosophy* 53(5). 177–187. https://doi.org/10.2307/2022451.

Reinhart, Tanya. 1990. Self-representation. Unpublished manuscript. http://www.let.uu.nl/~tanya.reinhart/personal/Papers/pdf/De_se_91.wp.pdf.

Santorio, Paolo. 2014. On the plurality of indices. Manuscript, University of Leeds. http://paolosantorio.net/opi.draft3b.pdf.

Schlenker, Philippe. 1999. *Propositional attitudes and indexicality: A cross categorial approach*. Massachusetts Institute of Technology dissertation. http:

//hdl.handle.net/1721.1/9353.
Schlenker, Philippe. 2003. A plea for monsters. *Linguistics and Philosophy* 26(1). 29–120. https://doi.org/10.1023/A:1022225203544.
Sharvit, Yael. 2010. Covaluation and unexpected BT effects. *Journal of Semantics* 28(1). 55–106. https://doi.org/10.1093/jos/ffq012.
von Stechow, Arnim. 2002. Binding by verbs: Tense, person and mood under attitudes. Unpublished manuscript. http://philpapers.org/rec/VONBBV.
von Stechow, Arnim. 2003. Feature deletion under semantic binding: Tense, person, and mood under verbal quantifiers. In *Proceedings of NELS*, vol. 33, 379–404.
Stephenson, Tamina. 2007. *Towards a theory of subjective meaning*. Massachusetts Institute of Technology dissertation. http://hdl.handle.net/1721.1/41695.
Stephenson, Tamina. 2010. Control in centred worlds. *Journal of semantics* 27(4). 409–436. https://doi.org/10.1093/jos/ffq011.
Yanovich, Igor. 2011. The problem of counterfactual de re attitudes. *Semantics and Linguistic Theory* 21. 56–75. https://doi.org/10.3765/salt.v21i0.2620.

Hazel Pearson
School of Languages, Linguistics and Film
Queen Mary University of London, London
Mile End Road
London, E1 4NS
United Kingdom
h.pearson@qmul.ac.uk